[Revised]

A Ph.D. SYNOPSIS

Research Area: **Computers and Soft Computing**

On the topic of

Soft Computing Techniques for improving Information Retrieval System

Submitted by

YOGESH GUPTA

DEPARTMENT OF ELECTRICAL ENGINEERING

FACULTY OF ENGINEERING

DAYALBAGH EDUCATIONAL INSTITUTE

(DEEMED UNIVERSITY)

AGRA-282110

(2014)

[Revised]

A Ph.D. SYNOPSIS

Research Area: **Computers and Soft Computing**

On the topic of

Soft Computing Techniques for improving Information Retrieval System

Submitted by

YOGESH GUPTA

Under the supervision of

Supervisor

Dr. Ashish Saini
Department of Electrical Engineering
Faculty of Engineering

Co-Supervisor

Prof. A.K. Saxena
Department of Electrical Engineering
Faculty of Engineering

Prof. V. Prem Pyara
Dean & Head (Electrical Engineering Dept.)
Faculty of Engineering

DAYALBAGH EDUCATIONAL INSTITUTE

(DEEMED UNIVERSITY)

AGRA-282110

(2014)

Soft Computing Techniques for improving Information Retrieval System

1. Introduction

The need to store and retrieve written information has become increasingly important over centuries, especially with the invention of paper and the printing press. The first systematic solution to the problem of finding the desired information from a large information collection was developed about 2,000 years ago by librarians, who kept track of "books" by cataloging them by author and the title. Searching through the catalog to find a book was a marked improvement from the physical search of actual books, but it required the searcher to know the book as well as to know its author and the title.

After the invention of computers, very soon people realized that they could be used for storing and mechanically retrieving large amounts of information. In 1945 Vannevar **[Van45]** published a ground breaking article titled "As We May Think" that gave birth to the idea of automatic access to large amounts of stored knowledge. In 1950s, this idea materialized into more concrete descriptions of how archives of text could be searched automatically. Several works emerged in mid 1950s that elaborated upon the basic idea of searching text with a computer. One of the most influential methods was described by H.P. Luhn **[Luh57]** in 1957, in which he proposed the use of words as indexing units for documents and for measuring word overlap as a criterion for retrieval.

Several key developments in this field happened in 1960s. Most notable were the development of the SMART system by Gerard Salton and his students **[Ger71]**, first at Harvard University and the Cranfield evaluations done by Cyril Cleverdon and his group **[Cle67]** at the College of Aeronautics in Cranfield. The Cranfield tests developed an evaluation methodology for retrieval systems that is still used even today by Information Retrieval (IR) systems. The SMART system, on the other hand, allowed researchers to experiment with ideas to improve search quality. A

system for experimentation coupled with good evaluation methodology allowed rapid progress in the field, and paved way for many critical developments. The basic idea was that people could find the desired information by selecting the "appropriate" keyword entry in the index, enlisting the documents related to it. Though this keyword index strategy expanded the capability of the catalog method by allowing the searcher to find a set of items related to a given concept (i.e., keyword), it introduced the problem of ambiguity in representation. Since there is no explicit rule for assigning keywords to documents, the choice of keywords for a given document depends heavily on the subjective word choice for a particular interpretation of the document. The obvious problem with this approach is the difficulty to select the "appropriate" keywords to express the information need.

The period of 1970s and 1980s saw many developments built on the advances of the 1960s. Various models for doing document retrieval were developed and advances were made along all dimensions of the retrieval process. These new models/techniques were experimentally proven to be effective on small text collections (several thousand articles) available to researchers at that time. However, due to lack of availability of large text collections, the question whether these models and techniques would scale to larger corpora remained unanswered. This changed in 1992 with the inception of Text REtrieval Conference (TREC) **[Har93]**. TREC is a series of evaluation conferences sponsored by various US Government agencies under the auspices of NIST, which aims at encouraging research in IR from large text collections. With large text collections available under TREC, many old techniques were modified, and many new techniques were developed (and are still being developed) to do effective retrieval over large collections. Now there are many large text collections like CACM, CISI, ADI etc.

1.1 Information Retrieval

The history of Information Retrieval parallels the development of libraries. The first civilizations had already come to the conclusions that efficient techniques should be designed to fully benefit from large document archives. Only recently, IR has radically changed with the advent of

computers. Digital technologies provide a unified infrastructure to store, exchange and automatically process large document collections.

Mooers **[Moo50]** and Savino **[Sav98]** have defined information retrieval as follows:

"Information retrieval is the name of the process or method whereby a prospective user of information is able to convert his need for information into an actual list of citations to documents in storage containing information useful to him."

Information retrieval system consists of two parts. First is the textual archive, which is a set of textual units (often called document collection), and second is a retrieval system with query. A user of a retrieval system presents queries describing what kinds of documents are desired. The retrieval system matches the queries against the documents in the textual archive. It then returns the user a list of sub-collection of the documents which are deemed as "best matches".

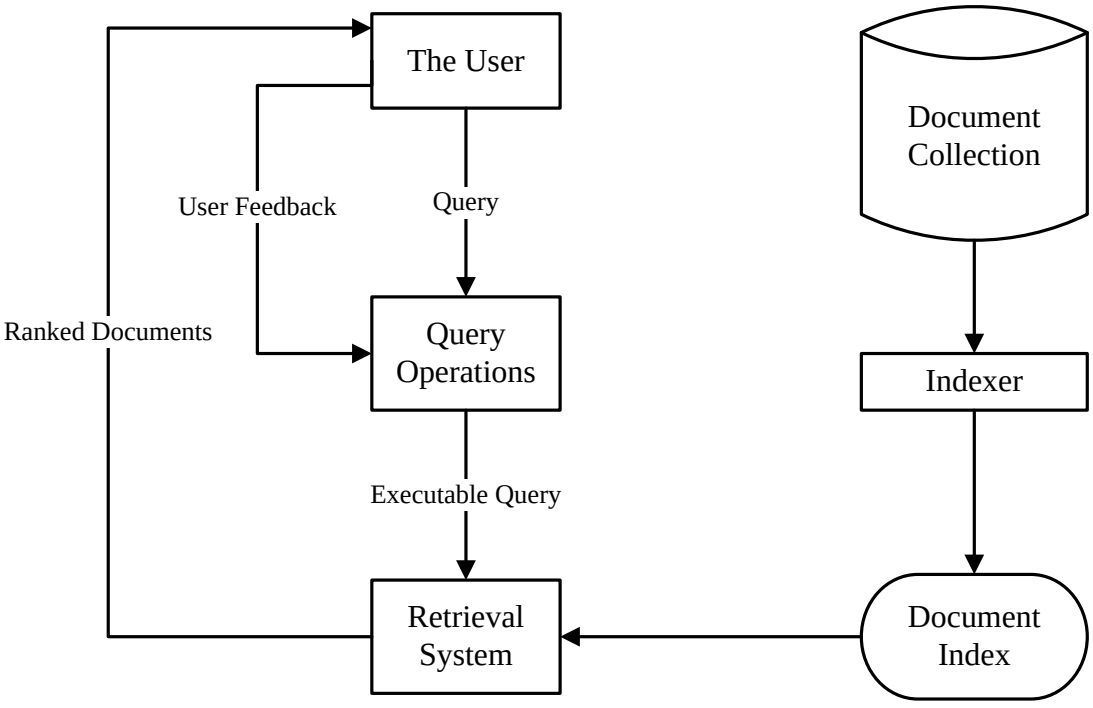

Fig.1. A general IR system architecture

A general information retrieval system architecture is shown In fig. 1 **[Liu07]**. In this figure, the user who needs information issues a query (**user query**) to the **retrieval system** through the **query operations** module. The retrieval module uses the **document index** to retrieve those documents that contain some query terms (such documents are likely to be relevant to the query), compute relevance scores for them, and then rank the retrieved documents according to the scores. The ranked documents are then presented to the user. The **document collection** is also called the **text database**, which is indexed by the **indexer** for efficient retrieval.

2. Important aspects of Information Retrieval

Although there are many aspects in Information Retrieval System but some prime aspects are document representation, similarity measure and query expansion.

2.1 Document Representation

Traditionally, documents may be available in different forms e.g. full text, hypertext, administrative text, directory, numeric or bibliographic text. It is very difficult to extract relevant information from these forms of documents. Therefore, first these documents should be represented in an appropriate manner with the help of any IR model. Such IR model provides the fundamental premises and forms the basis for ranking. In general, IR models operate on large and fixed collections of documents (corpus), from which they attempt to find out the useful information that best matches (are most relevant) to a user's need (query).

Baeza-Yates **[Yat99]** gives general definition of an IR model as:

Definition. *An IR model is a quadruple [D, Q, F, R* (q_i, d_j)*], where*

***1. D** is a set composed of logical views for the documents in the collection*
***2. Q** is a set composed of logical views for the user information needs expressed as queries*
***3. F** is a framework for modeling document representations, queries and their relationships*

4. R **(q_i, d_j)** *is a ranking function which associates a real number with a query $q_i \epsilon Q$ and a document representation $d_j \epsilon D$. Such ranking defines an ordering among the documents with regard to the query q_i.*

Bing Liu **[Liu07]** has presented various IR models. The main IR models are described as follows:

2.1.1 Boolean Model

The Boolean model **[Coo88]** was the first model which was adopted by most of the earlier systems and even today some of the commercial systems use this model, which makes use of the concepts of Boolean logic and set theories. The documents and the queries are a collection of terms and each term from the document is indexed. The presence and absence of a term in a document is represented by 1 and 0 respectively. For the term matching of document and query we maintain an inverted index of the terms i.e. for each term we must store a list of documents that contain the term. The terms are tokenized using linguistic models for those terms which can be stemmed down. The sequence of terms can be identified as < term, document ID> which can be sorted too. We can also have another identifier like frequency. Each query term specifies a set of documents containing the term and the Boolean operations performed on them are AND, OR and NOT.

Further, the Boolean model often retrieved either too many or too few documents due to the sensitive nature of the Boolean logic that responds rigidly to the absence or presence of a single term. To overcome the problem of output overload i.e., too many documents are retrieved without regard to their degree of potential importance to the user refinements to the system were made to produce ranked outputs by assigning weights to terms based on their "presumed" importance. Other refinement strategies, such as controlling the query formulation process to ease the difficulty of constructing complex Boolean queries, were investigated as well. While some tried to overcome the weaknesses of the Boolean model by building refinements to the existing Boolean model, others approached IR with a different search strategy called the Vector Space model.

2.1.2 Vector Space Model

The Vector Space Model, as the name implies, represents documents and queries internally in the form of vectors. In the vector space model all queries and documents are represented as vectors in |V|-dimensional space, where V is the set of all distinct terms in the collection (the vocabulary). A document vector contains index terms from the documents that to some extent describe its contents **[Sal98]**. At the center of the vector space model is the similarity measure, which is used to measure the angle between two vectors. The framework of the vector space model **[Wit99]** employs a ranking algorithm that tries to rank documents in order of how much of an overlap is between the terminology of the query and each document, where relatively rare terms have comparatively higher weights. Conceptually, documents are ranked on the basis of similarity measure. Some of the advantages of the Vector Space Model are that it is simple and fast model, that it can handle weighted terms, that it produces a ranked list as output and that the indexing process is automated which means a significantly lighter workload for the administrator of the collection. Also, it is easy to modify individual vectors, which is essential for the query expansion technique **[Sal98]**. The Vector Space Model has few weaknesses. The first weakness is the assumed independency between terms. Due to the locality of many term dependencies, their indiscriminate application to all the documents in the collection might badly affect the retrieval performance **[Yat99]**. Moreover, syntactic information remains unconsidered. The second weakness is that there are no theoretical justifications to use which similarity coefficients for a particular application and also some of the vector-manipulating operations.

The vector space model continues to be used in a variety of information retrieval areas apart from document retrieval, such as document categorization **[Joa97] [Hul94]** collaborative filtering **[Sob00]** or topic tracking **[Con04].**

2.1.3 Probabilistic Model

The Probabilistic model is similar to vector space model in its representation of documents and queries as vectors, but instead of retrieving documents based on their similarities to the query, the probabilistic model retrieves documents based on their probability of relevance to the query.

Rooted in the probabilistic notions introduced by **[Mar60]**, the probabilistic model views the principal function of IR as ranking of documents in the order of decreasing probability of relevance to a user's information need **[Rob77]**. The basic idea of the probabilistic model is to calculate the term weights, which define the probability of relevance of documents, based on the data about the distribution of query terms in documents that have been assessed for relevance. When term independence is assumed, the probability of relevance for a given document can be calculated by summing its individual term relevance weights, which are the estimations of probabilities that given terms in a query will appear in a relevant document but not in a non-relevant document. The probabilistic model suffers from the same limitation as the vector space model owing to the term independence assumption, an assumption introduced merely for the sake of computational simplicity.

2.2 Similarity Measures

The IR system needs to calculate the similarity of the query and the particular document in order to decide relevancy of that document with the query. When a document retrieval system is used to query a collection of documents with t terms, the system computes a vector D ($d_{i1}, d_{i2}, \ldots\ldots\ldots d_{it}$) of size t for each document. The vectors are filled with the weights and similarly, a vector Q ($W_{q1}, W_{q2}, \ldots\ldots\ldots W_{qt}$) is constructed for the terms found in the query. There are several typical vector similarity measures, such as the Inner product, the Dice coefficient, the Cosine coefficient, the Jaccard coefficient **[Sal98]** which can be used for finding similarity between query and the document.

1. Inner product: - The simplest similarity measure, the Inner product between a query Q and a document D_i, is defined by the product of the two vectors.

$$\textit{Inner } (Q, D_i) = \sum_{j=1}^{t} W_{qj} \; d_{ij} \qquad \textbf{(1)}$$

2. Cosine: - One drawback with using the Inner product is that longer documents, having more terms, will dominate the similarity calculations. Therefore, the vectors need to be normalized.

The most common of these is the cosine measure where the cosine of the angle between the query and document vector is given

$$Cos\ (Q, D_{i)} = \frac{\sum_{j=1}^{t} W_{qj}\ d_{ij}}{\sqrt{\sum_{j=1}^{t}(d_{ij})^2\ \sum_{j=1}^{t}(w_{qi})^2}} \quad \textbf{(2)}$$

The numerator represents the dot product (also known as the inner product) of the vectors q and d, while the denominator is the product of their Euclidean lengths. The effect of the denominator is thus to length-normalize the vectors. As the angle between the vectors shortens, the cosine angle approaches 1, meaning that the two vectors are getting closer, meaning that the vectors represent the similarity of document and query increases.

3. Dice Coefficient: - For document and query vector, the dice coefficient may be defined as twice the shared information (intersection) over the combined set (union)

$$Dice\ (Q, D_i) = \frac{2\sum_{j=1}^{t} W_{qj}\ d_{ij}}{\sqrt{\sum_{j=1}^{t}(d_{ij})^2\ \sum_{j=1}^{t}(w_{qi})^2}} \quad \textbf{(3)}$$

4. Jaccard Coefficient: - The Jaccard coefficient is defined as the size of the intersection divided by the size of the union of the document and query vectors

$$Jaccard\ (Q, D_i) = \frac{\sum_{j=1}^{t} W_{qj}\ d_{ij}}{\sum_{j=1}^{t}(d_{ij})^2 + \sum_{j=1}^{t}(w_{qi})^2 - \sum_{j=1}^{t} W_{qi}\, d_{ij}} \quad \textbf{(4)}$$

5. Okapi: - Okapi similarity measurement is one of the most popular methods used in the traditional IR field. Unlike VSM, the Okapi method not only considers the frequency of the query terms, but also the average length of the whole collection and the length of the document under evaluation.

$$\textbf{\textit{Okapi}}\ (Q, D_i) = \sum_{T \in Q} W \frac{(k_1+1)tf}{K+tf} \times \frac{(k_3+1)qtf}{k_3+qtf} \quad \textbf{(5)}$$

Q is a query that contains the words T

k1, b, and k3 are constant parameters (k1=1.2 and b=0.75 work well, k3 is 7 or 1000)

K is k_1 ((1-b) + (b · dl / avdl))

tf is the term frequency of the term with a document

qtf is the term frequency in the query

w is $\log \frac{(N-n+0.5)}{(n+0.5)}$

N is the number of documents, n is the number containing the term

dl and avdl are the document length and average document length

Okapi ranking uses the number of times a word occurs in a document, the number of documents containing the term, and the document length.

2.3 Query Expansion

Query Expansion is one of the promising approaches to deal with word mismatch problem in information retrieval **[Xu96]**. The basic idea of query expansion is to expand a user query by adding terms that are relevant to the original query terms. Since the expanded query contains more terms, the probability of matching them with terms in relevant documents is therefore increased. Three common types of query expansion are manual, interactive and automatic based on the role of involvement of user in whole process. One argument in favour of Automatic Query Expansion (AQE) is the system that has access to more statistical information on the relative utility of expansion terms and can make a better selection of which terms to add to the user's query.

Further distinction can divide AQE into two methods including global analysis and local feedback. The global analysis method relies on a thesaurus, typically constructed from a document corpus. Using the thesaurus, the global analysis method generates a ranked list of terms with respect to the original query terms and the top *n* terms are added to original query

[Xu96] [Qiu93] . On the other hand, the local feedback method first retrieves *N* documents that are most relevant to the original query, extracts the most important *n* terms from those documents, and subsequently, adds the extracted terns to the original query **[Xu96]**. As it assumes that top *N* documents are most relevant, it is also known as pseudo relevance feedback method. One of the problems inherent to the global analysis method for query expansion is that a global thesaurus is constructed and employed for expanding user queries. That is, a single weight for each pair of terms is derived from a collection of documents. Typically, the collection contains documents with different themes. For example, a set of information technology related documents may be classified into such themes as databases, artificial intelligence, computer architecture, and operating systems. In this case, the global view of term associations taken by the global analysis method for query expansion may not be adequate since the strengths/weights between two terms may be dissimilar or even totally different across different themes. For example, the terms "Feasibility Study" and "Quality Assurance" may be highly relevant under the theme of software engineering, while they are less relevant or even irrelevant in such themes as database and computer architecture. Thus, with the thematic view of term associations, for a user query "Feasibility Study," the term "Quality Assurance" should be added to original query under the theme of software engineering but should not be added under the theme of computer architecture. When taking the global view of term associations, the global weight between a pair of terms is a compromise of local weights across different themes. Thus, when expanding terms for a query, the global analysis method may select terms that are compromised across different themes rather than highly relevant terms in some of the themes in the document collection; thus, potentially limiting its retrieval effectiveness. In contrast to global analysis method, the local feedback method does not depend on a pre-constructed thesaurus for query expansion. Hence, it does not encounter the same problem as the global analysis does. Moreover, the local feedback method could result in better retrieval if the top *N* documents initially retrieved and used for feedback are in fact relevant to the original query **[Xu97]**.

2.4 Evaluation of performance of Information Retrieval System

The performance of any IR system can be evaluated by following four parameters.

1) **Precision:** Precision is a fraction of documents that are relevant among the entire retrieved document.
2) **Recall:** Recall is a fraction of the documents that are retrieved and relevant among all relevant documents.
3) **Precision-Recall Curve:** This curve is based upon the value of precision and recall where the x-axis is recall and y-axis is precision. Instead of using precision and recall on at each rank position , the curve is commonly plotted using 11 standard recall level 0%, 10%, 20% ……..100%.
4) **F-score:** F-score is harmonic mean of precision and recall.

3. Application of Soft Computing Techniques in IR

Soft computing is an emerging collection of methodologies, which aim to exploit tolerance for imprecision, uncertainty and partial truth to achieve robustness, tractability and total low cost. Soft computing methodologies have been advantageous in many applications. In contrast to analytical methods, soft computing methodologies mimic consciousness and cognition in several important respects: they can learn from experience; they can universalize into domains where direct experience is absent; and, through parallel computer architectures that simulate biological processes, they can perform mapping from inputs to the outputs faster than inherently serial analytical representations.

Soft-computing is a collection of techniques spanning many fields that fall under various categories in computational intelligence. Soft Computing has three main branches: fuzzy Systems, evolutionary computation, artificial neural computing, with the latter subsuming machine learning (ML) and probabilistic reasoning (PR), belief networks, chaos theory, parts of learning theory and wisdom based expert system (WES), etc. Fuzzy System is considered as a

soft computing technique because of realization of uncertainty, vagueness and ambiguity in real world problems.

As we know that user information needs are vague or imprecise and not easy to express in a question in Natural Language (NL). Sometimes user may change his query during information retrieval process and/or he may not be conscious of his exact needs of information. Therefore, to handle this uncertainty, vagueness and impreciseness, Fuzzy System is very suitable. It can be used for query term weighting and document clustering also.

Evolutionary Algorithm (EA) is a computational model based on natural evolution the whole process, which is highly random in nature. The EA process leads to filter individuals in the population closure to satisfying the objective functions of the optimization problem. EA have also all the characteristics of Soft Computing as it is highly robust and impression tolerance.

EA can be used for automatic document indexing **[Gor98]** to find relevant documents, matching function adaption **[Pat00]**, query optimization **[Che98] [Hor00]**, context based search **[Bau01]** etc.

4. Literature Review

4.1 Similarity Measure

Philip Resnik et al., **[Res95] [Res99]** presented measure of semantic similarity measure in an is-a taxonomy, based on the notion of information content in the year 1995 and 1999 respectively.

Jiang et al., **[Jia97]** combined a lexical taxonomy structure with corpus statistical information so that the semantic distance between nodes in the semantic space constructed by the taxonomy can be better quantified with the computational evidence derived from a distributional analysis of corpus data.

In year 1998, Dekang Lin et al., **[Lin98a]** proposed that, bootstrapping semantics from text is one of the greatest challenges in natural language learning. They defined a word similarity measure based on the distributional pattern of words. Dekang Lin et al., **[Lin98b]** presented an information theoretic definition of similarity that was applicable as long as there was a probabilistic model.

Fan presented similarity functions as trees and a classical generational scheme in **[Fan99]**.

In **[Fan00],** W. Fan presented a different approach to compute similarity measure to improve IR process. In **[Pat00]** Pathak et al. have proposed the idea of combined similarity measure in which they have proposed a linear combination of various similarity measures and then optimize the weight of each similarity measure using GA.

In 2004, Jian Pei et al., **[Pei04]** proposed a projection-based, sequential pattern growth approach for efficient mining of sequential patterns and Ming Li et al., **[Min04]** proposed a metric based on the non-computable notion of Kolmogorov computable distance and called it the similarity metric.

Mehran Sahami **[Sah06]** proposed a novel method for measuring the similarity between short text snippets by leveraging web search results to provide greater context for the short texts. In this paper, a method for measuring the similarity between short text snippets was proposed that captures more of the semantic context of the snippets rather than simply measuring their term-wise similarity. In the same year, Hsin-Hsi Chen **[Che06]** proposed a web search with double checking model to explore the web as a live corpus. Instead of simple web page counts and complex web page collection, the proposed novel model was a Web Search with Double Checking (WSDC) used to analyze snippets.

In 2007, Rudi L. Cilibrasi et al., **[Cil07]** proposed the words and phrases acquire meaning from the way they are used in society, from their relative semantics to other words and phrases. It was a new theory of similarity between words and phrases based on information distance and Kolmogorov complexity. The method was applicable to all search engines and databases.

Authors were introduced some notions underpinning the approach: Kolmogorov complexity, information distance, and compression-based similarity metric and a technical description of the Google distribution and the Normalized Google Distance (NGD). Hughes et al., **[Hug07]** proposed a method that presented the application of random walk Markov chain theory for measuring lexical semantic relatedness. Vincent Schickel-Zuber et al., **[Sch07]** presented a novel approach that allowed similarities to be asymmetric while still using only information contained in the structure of the ontology. Tuomo et al. used a connection between the cosine measure and the Euclidean distance in association with principal component analysis and grounded searching on the latter then applied the single and complete linkage and Ward clustering to Finnish documents utilizing their relevance assessment as a new feature in **[Tuo07].**

Ann Gledson et al., **[Ann08]** described a simple web-based similarity measure which relies on page counts only, could be utilized to measure the similarity of entire sets of words in addition to word pairs and could use any web-service enabled search engine distributional similarity measure. Torra et al. presented a method to calculate similarity between words based on dictionaries using Fuzzy graphs in **[Tor08]**.

In 2011, Bollegala et al., **[Bol11]** proposed a method which exploits the page counts and text snippets returned by a Web search engine. Chen presented a new similarity measure based on the geometric mean averaging operator to handle the similarity problems of generalized fuzzy numbers in **[Che11]**.

Usharani et al. **[Ush13]** proposed a genetic algorithm based method for finding similarity of web document based on cosine similarity.

4.2 Query Expansion

Query expansion is one of the important research topics of information retrieval systems. In order to improve the performance of information retrieval systems, some query expansion techniques have been proposed**.**

Van Rijsbergen proposed a relevance feedback technique to modify the original query by adding some other relevant terms in **[Van79].**

Yang and Korfaghe proposed a similar genetic algorithm to that of Robertson and Willet's in **[Yan94]**. They used a real coding, and the two-point crossover and random mutation operators (besides, crossover and mutation probabilities were changed throughout the algorithm run). The selection was based on a classic generational scheme where the chromosomes with a fitness value below the average of the population were eliminated and the reproduction was performed by Baker's mechanism. The results were satisfactory. However, while the average precision for some levels of recall increased after this process, the same behavior was not noticed in the average recall for some levels of precision.

In **[San95]**, Sanchez et al. proposed a genetic algorithm to learn the term weights of extended Boolean queries for fuzzy Information Retrieval systems in a relevance feedback process. Binary-coded chromosomes encoded the n term weights as well as the similarity threshold considered in the document retrieval (stored in the n $+1^{th}$ gene). The genetic operators were the classic ones and the fitness function was based on a linear combination of precision and recall. The behavior of the system was studied on a 479 document collection about patents. Unfortunately, they did not show the obtained results in the paper.

Robertson and Willet **[Rob96]** proposed a genetic algorithm to investigate an upper bound for relevance feedback techniques for query expansion in vector space information retrieval systems and compared its results with Robertson and Spark Jones's **[Rob76]** retrospective relevance weights technique. This technique did not include negative relevance weights.

Xu and Croft used the local analysis and the global analysis of documents for query expansion in **[Xu96]**.

Kraft et al. **[Kra97]** proposed a query expansion technique to learn the whole composition of extended Boolean queries for Fuzzy IR systems. The algorithm is based on genetic programming. The preliminary results indicated that randomly selecting terms from the set of all

terms to populate queries did not work efficiently. To solve this drawback the terms were selected from the predetermined documents specified as relevant. It obtained good results but it suffered from one of the main limitations of the genetic programming paradigm: the learning of the weights considered in the encoded structure could only be performed by mutation.

Cooper and Byrd constructed a visual interface with graphical relations between items by lexical neighborhoods for prompted query refinement in **[Coo98]**. In **[Che98],** Chen et al. used a GA as an IQBE technique to learn query terms that better represent a relevant document set provided by the user.

In **[Hor00]** the author has used a GA to adapt the query term weights in order to get the closest query vector to the optimal one. Li and Agrawal used multi-granularity indexing and query processing for supporting the web query expansion in **[Agr00]** and in **[Wei00]** Wei et al. presented a method to mine term association rules for automatic global query expansion.

Chen et al. used association rules to discover the degrees of similarity between terms and constructed a hierarchical-tree structure to pick out query expansion terms in **[Che01]**. In the same year Takagi and Tajima presented a method for query expansion using conceptual fuzzy sets for search engines in **[Tak01]**. It calculates the degrees of similarity between terms to construct a hierarchical tree structure and lets terms with higher degrees of similarity be expansion terms of the structure. and in **[Kim01]** Kim et al. also presented a method for query term expansion and reweighting using the term co-occurrence similarity and fuzzy inference techniques.

Cui et al. presented a method for probabilistic query expansion using query logs in **[Cui02]**.

Billerbeck et al. proposed a method for query expansion using associated queries in **[Bil03]**. In **[Cha03]** Chang et al. presented a query expansion method based on fuzzy rules. In the same year Jin et al. developed a method for query expansion based on the term similarity tree model in **[Jin03]**. In **[Lat03a] [Lat03b]** Latiri et al. considered the relationship between terms and documents as a fuzzy binary relation, based on the closure of the extended fuzzy Galois

connection, and used fuzzy association rules to find out real correlated terms as query expansion terms. Nakauchi et al. created thesaurus and relationships of terms for query expansion in **[Nak03]**. In the same year Safar and Kefi presented a query expansion method based on the domain ontology and the lattice structure in **[Saf03]**.

Berardi et al. used association rules to mine query expansion terms and presented how to filter off redundant association rules in **[Ber04]**. Martin-Bautista et al. presented a method to mine web documents for finding additional query terms in **[Mar04]**. Stojanovic used a conceptual schema to query neighborhood for query expansion in **[Sto04]**.

Lin et al. presented a method for mining additional query terms for query expansion in **[Lin05].** In **[Bei05],** Michel Beigbeder and Annabelle Mercier proposed a IR model using the fuzzy proximity degree of term occurrences.

Chang et al. presented a new method for query reweighting to deal with document retrieval in **[Cha06].** Grootjen et al. presented a new, hybrid approach that projects an initial query result onto global information, yielding a local conceptual overview in **[Gro06].** In **[Bil06]**, Billerbeck et al. proposed a new method that draws candidate terms from brief document summaries that are held in memory for each document. In **[Lin06]**, Lin et.al. proposed a method for query expansion based on user relevance feedback techniques for mining additional query terms. According to the user's relevance feedback, the proposed query expansion method calculates the degree of importance of relevant terms of documents in the document database.

Chang et al. proposed a new query expansion method for document retrieval based on fuzzy rules in **[Cha07].**

Nowacka et al. proposed a comprehensive fuzzy based model of information retrieval in **[Now08].** Fattahi et al presented a new approach to query expansion in search engines through the use of general non-topical terms (NTTs) and domain-specific semi-topical terms (STTs) in **[Fat08].** Cecchini et al. proposed techniques place emphasis on searching for novel material that is related to the search context in **[Cec08].**

Carlos et al. proposed a semi-supervised algorithm to incrementally learn terms that can help bridge the terminology gap existing between the user's information needs and the relevant documents' vocabulary in **[Car09].**

Piotr Wasilewski proposed a method for query expansion using semantic modeling of information need in **[Pio11].** Liu et al. in **[Liu11]** proposed two algorithms for query expansions. First is iterative single keyword refinement and second is elimination based convergence.

Tayal et al. presented a method for fuzzy weighting of query terms with the help of fuzzy triangular membership function in **[Tay12]**. Latiri et al. proposed automatic query expansion method using association rule mining approach in **[Lat12]**.

5. Research Gap and Motivation

Nowadays, automatic information retrieval systems are widely used in several application domains (e.g. web search, digital library search, blog search, information filtering, recommender system and social search etc.) and there is a constant need to improve such systems. In this context, Information Retrieval is an active field of research within Computer Science. The major concern of IR is to find the documents relevant to submitted queries. Similarity measures and query expansion play important roles in this area.

Although some efforts have been made to create IR systems that can index billions of documents, but the studies of such IR systems behavior have shown that a large portion of queries crafted by user consist of only one to three terms. It is difficult to achieve high degree of relevancy of documents in information retrieval by using such short queries crafted by user. Therefore, there is need of extensive research to explore query expansion approaches with respect to their efficiency in improving retrieval effectiveness and to find more efficient similarity measure using soft computing techniques like fuzzy system and evolutionary algorithms etc. On the basis of literature survey, the following research gap can be identified:

- Conventional statistical similarity measures fail to capture inherent features of documents & queries due to subjectivity involved in natural language text. Here, fuzzy logic based similarity measure can be developed to address uncertainty, vagueness and to capture untouched features in documents and queries. Fuzzy logic provides a convenient way of converting existing knowledge into fuzzy logic rules.
- Although few researchers like Pathak et.al. **[Pat00]** have proposed hybrid similarity measure but it was linear combination of statistical measures where combination of weights were determined for maximum fitness value based on *precision* only.
- Mostly query expansion techniques use statistical method (such as TF.IDF method) to assign weight to query terms, but best values of weights cannot be determined. It is a problem of optimization, so genetic algorithm can be used to weight and to reweight the query terms.
- Queries are written in natural languages, which include uncertainty and vagueness. Therefore, fuzzy logic can be used to choose appropriate candidate terms for query expansion.
- Due to the well known strengths of fuzzy logic, the investigation of proper amalgam of pseudo relevance feedback and fuzzy logic is still required.

6. Research Objectives

In this research work, we propose to take the following objectives:

- To develop an efficient *similarity measure* using fuzzy logic to improve the performance of IR System and to compare its performance with other similarity measures reported in literature.
- To propose new approaches for *automatic query expansion* using soft computing techniques in *term weighting* and *pseudo relevance feedback* so as to enhance the efficiency of IR process and analyze their performance.

References

[Van45] Vannevar Bush, *As We May Think,* Atlantic Monthly, Vol 176, pp 101–108, July 1945.

[Moo50] Mooers C. N., *Information retrieval viewed as temporal signaling,* Proceedings of the International Congress of Mathematicians, Vol 1, pp 572–573, 1950.

[Luh57] H. P. Luhn, *A statistical approach to mechanized encoding and searching of literary information,* IBM, Journal of Research and Development, 1957.

[Mar60] Maron M. E. and Kuhns J. L., *On relevance, probabilistic indexing and information Retrieval,* Journal of the Association for Computing Machinery, Vol 7, pp 216-244, 1960.

[Cle67] C.W. Cleverdon, *The Cranfield tests on index language devices,* Aslib Proceedings, 19, pp 173–192, 1967.

[Ger71] Gerard Salton, *The SMART Retrieval System—Experiments in Automatic Document Retrieval*, Prentice Hall Inc., Englewood Cliffs, NJ, 1971.

[Rob76] Robertson S.E. and Spark Jones, *Relevance weighting of search terms*, Journal of the American Society for Information Science, Vol. 27, pp 129-145, 1976.

[Rob77] Robertson S. E., *The probability ranking principle in IR,* Journal of Documentation, Vol 33, pp 294-304, 1977.

[Van79] C.J. Van Rijsbergen, *Information Retrieval*, second edition, Butterworth, USA, 1979.

[Coo88] Cooper W.S., *Getting beyond Boole,* Information Processing and Management, Vol 24, pp 243-225, 1988.

[Har93] D. K. Harman, *Overview of the first Text REtrieval Conference (TREC-1),* In Proceedings of the First Text REtrieval Conference (TREC-1), pp 1–20. NIST Special Publication 500-207, March 1993.

[Qiu93] Qiu Y. and Frei H. P., *Concept based query expansion,* In Proceedings of the 16th annual international ACM SIGIR conference on Research and development in information retrieval, SIGIR '93, pp 160-169, NY, USA, ACM Press, 1993.

[Hul94] Thomas C. Hull, *On the mathematics of flat origamis*, Congressus Numerantium, Vol 100, pp 215-224, 1994.

[Yan94] Yang J. and Korfhage R., *Query modifications using genetic algorithms in vector space models*, International Journal of Expert Systems, Vol. 7, No. 2, pp 165-191, 1994.

[Res95] Resnik P, *Using Information Content to Evaluate Semantic Similarity in a Taxonomy*, 14th International Joint Conference on Artificial Intelligence, Vol 1, pp 24-26, 1995.

[San95] Sanchez E., Miyano H. and Brachet J., *Optimization of fuzzy queries with genetic algorithms*. In proceedings of Applications to a data base of patents in biomedical engineering, VI IFSA Congress, Sao-Paulo, Brazil, pp 293-296, 1995.

[Rob96] Robertson A. and Willet P., *An upperbound to the performance for ranked-output searching: optimal weighting of query terms using a genetic algorithm*, Journal of Documentation Vol. 52, No. 4, pp 405-420, 1996.

[Xu96] Xu J. and Croft W. B., *Query expansion using local and global document analysis*, Proceedings of the 19th annual international ACM SIGIR conference on research and development in information retrieval, Zurich, Switzerland, pp 4–11, 1996.

[Joa97] Thorsten Joachims, *A Probabilistic Analysis of the Rocchio Algorithm with TF-IDF for Text Categorization*, In Proceedings of the Fourteenth International Conference on Machine Learning (ICML 1997), Nashville, Tennessee, USA, 1997.

[Kra97] Kraft D.H., Petry F.E., Buckes B.P. and Sadasivan T., *Genetic algorithm for query optimization in information retrieval: relevance feedback*, Genetic Algorithms and Fuzzy Logic Systems, pp 55-173, 1997.

[Xu97] Xu J., *Solving the word mismatch problem through text analysis*, Ph.D. Thesis, University of Massachusetts, Department of Computer Science, Amherst, USA, 1997.

[Che98] H. Chen et al., *A machine learning approach to inductive query by examples: an experiment using relevance feedback, ID3, genetic algorithms, and simulated annealing*, Journal of the American Society for Information Science, Vol 49, No 8, pp 693–705, 1998.

[Coo98] Cooper J. W. and Byrd R. J., *OBIWAN—a visual interface for prompted query refinement,* Proceedings of the 31st Hawaii international conference on system sciences, Hawaii, Vol 2, pp 277–285, 1998.

[Gor98] Gordan M., *Probabilistic and Genetic Algorithms for Document Retrieval,* Communication of ACM, Vol 31, No 120, pp 1208-1218, 1998.

[Lin98a] Lin D., *Automatic Retrieival and Clustering of Similar Words*, International Committee on Computational Linguistics and the Association for Computational Linguistics, pp 768-774, 1998.

[Lin98b] Lin D., *An Information-Theoretic Definition of Similarity*, 15th International Conference on Machine Learning, pp 296-304, 1998.

[Sal98] Salton G., *Automatic text processing: the transformation, analysis, and retrieval of information by computer,* Addison-Wesley, 1998.

[Sav98] Savino P. and Sebastiani F., *Essential bibliography on multimedia information retrieval, categorization and filtering,* In Slides of the 2nd European Digital Libraries Conference Tutorial on Multimedia Information, 1998.

[Fan99] W. Fan, M. Gordon and P. Pathak, *Automatic generation of a matching function by genetic programming for effective information retrieval,* in America's Conference on Information System, Milwaukee, USA, 1999.

[Res99] Resnik P, *Semantic Similarity in a Taxonomy: An Information based Measure and its Application to problems of Ambiguity in Natural Language*, Journal of Artificial Intelligence Research, Vol 11, pp 95-130, 1999.

[Wit99] Witten I., Moffat A. and Bell T., *Managing Gigabytes: Compressing and Indexing Documents and Images*, Morgan Kaufmann, 1999.

[Yat99] Yates R. B. and Berthier R., *Modern Information retrieval,* Addisson Wesley, 1999.

[Agr00] Li W. S. and Agrawal D., *Supporting web query expansion efficiently using multi-granularity indexing and query processing*, Journal of Data and Knowledge Engineering, Elsevier, Vol 35, No 3, pp 239–257, 2000.

[Fan00] W. Fan, M.D. Gordon and P. Pathak, *Personalization of search engine services for effective retrieval and knowledge management,* in Proceedings of International Conference on Information Systems (ICIS), Brisbane, Australia, 2000.

[Hor00] J. Horng and C. Yeh, *Applying genetic algorithms to query optimization in document retrieval*, Information Processing and Management, Elsevier, Vol 36, pp 737–759, 2000.

[Pat00] P. Pathak, M. Gordon and W. Fan, *Effective information retrieval using genetic algorithms based matching functions adaption*, in Proceedings of 33rd Hawaii International Conference on Science (HICS), Hawaii, USA, 2000.

[Sob00] Ian Soboroff and Charles Nicholas, *Collaborative Filtering and the Generalized Vector Space Model (Poster),* Proceedings of the 23rd Annual International Conference on Research and Development in Information Retrieval (SIGIR 2000), Athens, Greece, 2000.

[Wei00] Wei J., Bressan, S. and Ooi B. C., *Mining term association rules for automatic global query expansion: Methodology and preliminary results,* Proceedings of the first international conference on web information systems engineering, Hong Kong, China, Vol 1, pp 366–373, 2000.

[Bau01] Bauer T., and Leake, *WordSieve: A method for real-time context extraction in Modeling and Using Context*, in Proceedings of the Third International and Interdisciplinary Conference, Berlin, Springer-Verlag, 2001.

[Che01] Chen H., Yu J. X., Furuse K. and Ohbo N., *Support IR query refinement by partial keyword set*, Proceedings of the second international conference on web information systems engineering, Singapore, Vol 1, pp 245–253, 2001.

[Deb01] Kalyanmoy Deb, *Multi-Objective Optimization using Evolutionary Algorithms*, USA, John Wiley & Sons, Ltd., 2001.

[Kim01] Kim B. M., Kim J. Y. and Kim J., *Query term expansion and reweighting using term co-occurrence similarity and fuzzy inference*, Proceedings of the joint ninth IFSA world congress and 20th NAFIPS international conference, Vancouver, Canada, Vol 2, pp 715–720, 2001.

[Laf01] Lafferty J. and Zhai C., *Document language models, query models, and risk minimization for information retrieval*. In SIGIR '01, Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval, New York, USA, ACM Press, pp 111-119, 2001.

[Sak01] Sakai T. and Robertson S. E., *Flexible pseudo-relevance feedback using optimization tables*, Louisiana, pp 396-397, 2001.

[Spi01] Amanda Spink, Dietmar Wolfram, B. J. Jansen, and Tefko Saracevic, *"Searching the Web: The public and their queries"* Journal of the American Society for Information Science and Technology, Vol 52, No 3, pp 226–234, 2001.

[Tak01] Takagi T. and Tajima M., *Query expansion using conceptual fuzzy sets for search engine*, Proceedings of the 10th IEEE international conference on fuzzy systems, Melbourne, Australia, pp 1303–1308, 2001.

[Zha01] Zhai C. and Lafferty J., *Model-based Feedback in the Language Modeling approach to Information Retrieval*, In CIKM '01, Proceedings of the 10th International Conference on Information and Knowledge Management, New York, USA, ACM Press, pp 403-410, 2001.

[Cui02] Cui H., Wen J. R., Nie J. Y. and Ma W. Y., *Probabilistic query expansion using query logs*, Proceedings of the 11th international conference on World Wide Web, Honolulu, Hawaii, pp 325–332, 2002.

[Kli02] Klink S., Hust A., Junker M. and Dengel, A., *Improving Document Retrieval by Automatic Query Expansion Using Collaborative Learning of Term-Based Concepts*, Document Analysis Systems, pp 376-387, 2002.

[Bil03] Billerbeck B., Scholer F., Williams H. E. and Zobel J., *Query expansion using associated queries*, Proceedings of the 12th international conference on information and knowledge management, New Orleans, LA, pp 2–9, 2003.

[Cha03] Chang Y. C., Chen S. M. and Liau C. J., *A new query expansion method based on fuzzy rules,* Proceedings of the seventh joint conference on AI, Fuzzy system, and Grey system, Taipei, Taiwan, Republic of China, 2003.

[Cui03] Cui H., Wen J. R., Nie J.Y. and Ma W.Y., *Query expansion by mining user logs,* Knowledge and Data Engineering, IEEE Transactions, Vol 15, No 4, pp 829-839, 2003.

[Jin03] Jin Q., Zhao J., and Xu B., *Query expansion based on term similarity tree model*, Proceedings of the 2003 international conference on natural language processing and knowledge engineering, Beijing, China, pp 400–406, 2003.

[Lat03a] Latiri C. C., Elloumi S., Chevallet J. P. and Jaoua A., *Extension of fuzzy Galois connection for information retrieval using a fuzzy quantifier,* Proceedings of the 2003 ACS/IEEE international conference on computer systems and applications, Tunis, Tunisia, 2003.

[Lat03b] Latiri C. C., Yahia S. B., Chevallet J. P. and Jaoua A., *Query expansion using fuzzy association rules between terms*, Proceedings of the 2003 fourth JIM international conference on knowledge discovery and discrete mathematics, Mets, France, 2003.

[Nak03] Nakauchi K., Ishikawa Y., Morikawa H. and Aoyama T., *Peer-to-peer keyword search using keyword relationship,* Proceedings of the third IEEE/ACM international symposium on cluster computing and the grid, Tokyo, Japan, pp 359–366, 2003.

[Saf03] Safar B. and Kefi H., *Domain ontology and Galois lattice structure for query refinement,* Proceedings of the 15th IEEE international conference on tools with artificial intelligence, Sacramento, California, pp 597–601, 2003.

[Ama04] Amati G., Carpineto C. and Romano G., *Query difficulty, robustness and selective application of query expansion*, In European Conference on Information Retrieval (ECIR), pp 127-137, 2004.

[Bac04] Bacchin M. and Melucci M., *Expanding Queries using Stems and Symbols*, In Proceedings of the 13th Text REtrieval Conference (TREC 2004) Genomics Track, Gaithersburg, MD, USA, 2004.

[Ber04] Berardi M., Lapi M., Leo P., Malerba D., Marinelli C. and Scioscia G., *A data mining approach to PubMed query refinement*, Proceedings of the 15th international workshop on database and expert systems applications, Zaragoza, Spain, pp 401–405, 2004.

[Con04] M. Connell, A. Feng, G. Kumaran, H. Raghavan, C. Shah and J. Allan, *Topic Detection and Tracking Workshop*, UMass at TDT, 2004.

[Mar04] Martin-Bautista M. J., Sanches D., Chamorro-Martinez J., Serrano J. M. and Vila M. A., *Mining web documents to find additional query terms using fuzzy association rules*, Fuzzy Sets and Systems, Elsevier, Vol 148, No 1, pp 85–104, 2004.

[Min04] Ming Li, Xin Chen, Xin Li, Bin Ma, Paul M. and B. Vitnyi, *The Similarity Metric*, IEEE Transactions on Information Theory, Vol 50, No 12, pp 3250-3264, 2004.

[Pei04] Pei J., Han J., Mortazavi Asi B., Wang J., Pinto H., Chen Q., Dayal U. and Hsu M., *Mining Sequential Patterns by Pattern growth: the Prefix span Approach*, IEEE Transactions on Knowledge and Data Engineering, Vol 16, No 11, pp 1424-1440, 2004.

[Sta04] Staff C. and Muscat R., *Expanding Query Terms in Context*, In Proceedings of Computer Science Annual Workshop (CSAW'04), University of Malta, pp 106-108, 2004.

[Sto04] Stojanovic N., *On using query neighborhood for better navigation through a product catalog: SMART approach*, Proceedings of the 2004 IEEE international conference on e-Technology, e-Commerce and e-Service, Taipei, Taiwan, Republic of China, pp 405–412, 2004.

[Bei05] Michel Beigbeder and Annabelle Mercier, *An Information Retrieval Model using the Fuzzy Proximity Degree of Term occurrences*, SAC'05, Santa De, New Mexico, USA, 2005.

[Col05] Collins-Thompson K. and Callan J., *Query expansion using random walk models,* In CIKM '05: Proceedings of the 14th ACM international conference on Information and knowledge management, New York, USA, ACM Press, pp 704-711, 2005.

[Gon05] Gong Z., Cheang C. and Leong Hou U., *Web query expansion by wordnet,* In Database and Expert Systems Applications, Vol 3588, Lecture Notes in Computer Science, Springer Berlin / Heidelberg, pp 166-175, 2005

[Lin05] Lin H. C., Wang L. H. and Chen S.M., *A new query expansion method for document retrieval by mining additional query terms*, Proceedings of the 2005 International conference on business and information, Hong Kong, China, 2005.

[Sak05] Sakai T., Manabe T. and Koyama M., *Flexible pseudo-relevance feedback via selective sampling,* ACM Transactions on Asian Language Information Processing (TALIP), Vol 4, No 2, pp 111-135, 2005.

[Wan05] Wang L. H., Lin H. C. and Chen S. M., *A new method for query expansion based on uses relevance feedback techniques*, Proceedings of the Sixth International Symposium on Advanced Intelligent Systems, Neosn Korea, pp 679–684, 2005.

[Bil06] Bodo Billerbeck and Justin Zobel, *Efficient query expansion with auxiliary data structures*, Information Systems 31, pp 573–584, 2006.

[Cha06] Yu-Chuan Chang and Shyi-Ming Chen, *A New Query Reweighting Method for Document Retrieval Based on Genetic Algorithms*, IEEE Transactions On Evolutionary Computation, Vol 10, No 5, pp 617-622, 2006.

[Che06] Chen H., Lin M. and Wei Y., *Novel Association Measures using Web Search with Double Checking*, International Committee on Computational Linguistics and the Association for Computational Linguistics, pp 1009-1016, 2006.

[Gro06] F.A. Grootjen and T.P. van der Weide, *Conceptual Query Expansion,* Data and Knowledge Engineering, Elsevier, Vol 56, pp 174–193, 2006.

[Lin06] Lin H.C., Wang L.H. and Chen S.M., *Query expansion for document retrieval based on fuzzy rules and user relevance feedback techniques*, Expert Systems with Applications, Vol. 31, pp 397-405, 2006.

[Liu06] Bing Liu. *Web Data Mining,* A Book Published by Springer, 2006.

[Sah06] Sahami M. and Heilman T., *A Web-based Kernel Function for Measuring the Similarity of Short Text Snippets*, 15th International Conference on World Wide Web, pp 377-386, 2006.

[Tao06] Tao T. and Zhai C., *Regularized estimation of mixture models for robust pseudo-relevance feedback*, In SIGIR '06: Proceedings of the 29th Annual International ACM SIGIR conference on Research and Development in Information Retrieval, New York, USA. ACM Press, pp 162-169, 2006.

[Voo06] Voorhees E., *Overview of the TREC 2005 robust retrieval track*, In E.M. Voorhees and L. P. Buckland, editors, The Fourteenth Text Retrieval Conference, TREC 2005, Gaithersburg, MD. NIST, 2006.

[Cha07] Yu Chuan Chang, Shyi Ming Chen and Churn Jung Liau, *A new query expansion method for document retrieval based on the inference of fuzzy rules*, Journal of Chinese Institute of Engineers, Vol 30, No 3, pp 511-515, 2007.

[Cil07] Cilibrasi R. and Vitanyi P., *The google similarity distance*, IEEE Transactions on Knowledge and Data Engineering, Vol 19, No 3, pp 370-383, 2007.

[Hug07] Hughes T. and Ramage D., *Lexical Semantic Relatedness with Random Graph Walks*, Conference on Empirical Methods in Natural Language Processing Conference on Computational Natural Language Learning, (EMNLP-CoNLL07), pp 581-589, 2007.

[Liu07] Bing Liu, *Web Data Mining: Exploring Hyperlinks, Contents and Usage Data*, Chicago, USA, Springer-Verlag, Berlin Heidelberg, 2007.

[Met07] Metzler D. and Croft W. B., *Latent concept expansion using markov random fields*, In SIGIR '07: Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval, New York, USA. ACM Press, pp 311-318, 2007.

[Sch07] Schickel Zuber V. and Faltings B, *OSS: A Semantic Similarity Function Based on Hierarchical Ontologies*, International Joint Conference on Artificial Intelligence, pp 551-556, 2007.

[Tuo07] Tuomo Korenius, Jorma Laurikkala and Martti Juhola, *On principal component analysis, cosine and Euclidean measures in information retrieval*, Information Science, Elsevier, Vol 177, pp 4893-4905, 2007.

[Ann08] Ann Gledson and John Keane, *Using Web-Search Results to Measure Word-Group Similarity*, 22nd International Conference on Computational Linguistics, pp 281-288, 2008.

[Cao08] Cao G., Nie J.Y., Gao J. and Robertson S., *Selecting good expansion terms for pseudo-relevance feedback*, In SIGIR '08: Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval, New York, USA, ACM Press, pp 243-250, 2008.

[Cec08] Rocı´o L. Cecchini, Carlos M. Lorenzetti, Ana G. Maguitman and Ne´lida Beatrı´z Brignole, *Using genetic algorithms to evolve a population of topical queries*, Information Processing and Management, Elsevier, Vol 44, pp 1863–1878, 2008.

[Fat08] Rahmatollah Fattahi, Concepcio´n S. Wilson and Fletcher Cole, *An alternative approach to natural language query expansion in search engines: Text analysis of non-topical terms in Web documents*, Information Processing and Management, Elsevier, Vol 44, pp 1503–1516, 2008.

[Man08] Manning C.D., Raghavan P. and Schtze H., *Relevance feedback and query expansion*, In Proceedings of Introduction to Information Retrieval, Cambridge University Press, New York, 2008.

[Now08] Nowacka K., Zadrozny S. and Kacprzyk J., *A new fuzzy logic based information retrieval model*, In Proceeding of IPMU'08, pp 1749-1756, 2008.

[Tor08] Vicenc Torra and Yasuo Narukawa, *Word Similarity from dictionaries: Inferring Fuzzy measures and Fuzzy graphs*, International Journal of Computational Intelligence Systems, Vol 1, No 1, pp 19–23, 2008.

[Car09] Carlos M. Lorenzetti and Ana G. Maguitman, *A semi-supervised incremental algorithm to automatically formulate topical queries*, Information Sciences, Elsevier, Vol 179, pp 1881–1892, 2009.

[Xu09] Xu Y., Jones G. J. and Wang B., *Query dependent pseudo-relevance feedback based on Wikipedia,* In SIGIR '09, Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval, New York, USA, ACM, pp 59-66, 2009.

[Yin09] Yin Z., Shokouhi M. and Craswell N., *Query expansion using external evidence*, In Boughanem M., Berrut C., Mothe J., and Soule-Dupuy C. editors, Advances in Information Retrieval, Vol 5478, Lecture Notes in Computer Science, Springer Berlin / Heidelberg, pp 362-374, 2009.

[Zha10] Lv Y. and Zhai C., *Positional relevance model for pseudo-relevance feedback*, In Proceeding of the 33rd international ACM SIGIR conference on Research and development in information retrieval, SIGIR '10, New York, USA, ACM, pp 579-586, 2010.

[Bol11] Danushka Bollegala, Yutaka Matsuo and Mitsuru Ishizuka, *A Web Search Engine-based Approach to Measure Semantic Similarity between Words*, IEEE Transactions on Knowledge and Data Engineering , Vol 23, No 7, pp 977-990, 2011

[Che11] Shi Jay Chen, *Fuzzy information retrieval based on a new similarity measure of generalized fuzzy numbers,* Intelligent Automation and Soft Computing, Vol 17, No 4, pp 465-476, 2011.

[Liu11] Ziyang Liu, Sivaramakrishnan Natarajan and Yi Chen, *Query Expansion based on Clustered Results*, Proceedings of the VLDB Endowment, Vol 4, No 6, 2011.

[Pio11] Piotr Wasilewski, *Query Expansion by Semantic Modeling of Information Need,* Proceedings of International Workshop CS&P, 2011.

[Lat12] Chiraz Latiri, Hatem Haddad and Tarek Hamrouni, *Towards an effective automatic query expansion process using an association rule mining approach*, Journal of Intelligent Information System, pp 209-247, 2012.

[Tay12] Devendra K. Tayal, Smita Sabharwal, Amita Jain and Kanika Mittal, *Intelligent query expansion for the queries including numerical terms*, National Conference on Communication Technologies and its impact on Next Generation Computing CTNGC 2012, Proceedings published by International Journal of Computer Applications, pp 35-39, 2012.

[Ush13] J. Usharani and K. Iyakutti, *A Genetic Algorithm based on Cosine Similarity for Relevant Document Retrieval,* International Journal of Engineering Research & Technology (IJERT), Vol 2, No 2, 2013.